School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 1, 2019)
Workshop exercises: Week 8

1. Complete any remaining questions from the previous week, if necessary.

2. **Zoning** is the process of adjusting the weights of tokens, depending on document metadata (for example, if a token appears in the page title, or an image caption, it might be assigned a different weight). Why is this a stage in the **parsing** step of an IR engine, and not the **indexing** or **querying** step?

   - In short, because the document metadata is stripped during the parsing step — we are left with a stream of tokens, and generally no indication about the information necessary to apply zoning (like where in the document each term appeared).
   - It's true that the actual weights are stored in the inverted index and (probably) calculated for the TF-IDF model at query time. However, parsing is about building a representation of each document in the collection; to account for the document metadata, we require an alternative parsing strategy.

3. What is a **phrase query**, and why is an inverted index — like the one from the question above — inadequate for phrase querying? How could the index be altered to support this style of querying?

   - Phrase querying means returning only documents where the query terms appear in sequence. Since the index above doesn't contain any information about term ordering (only term frequency), we can't assess whether the query appears in sequence or not.
   - The most common strategy for dealing with phrasal queries is to use a "positional index": an index where the positions of terms (indices within the document) are stored, as well as the term frequencies.

4. What is **link analysis**? What aspects of user behaviour or the nature of data on the Web is it trying to model?

   - Link analysis is a weighting (or re-ranking) procedure which alters the importance of documents (or terms within documents), based on the link structure of the Web.
   - For "popularity"–based approaches, we're trying to build a model of how popular pages are based on which pages are linking to a given document (and, in turn, the popularity of those pages). We then assume that popular pages are more likely to be relevant than unpopular pages, and the former are pushed up the ranking, while the latter are dropped down. PageRank is an example of one such algorithm.

5. In the **PageRank** algorithm:

   (a) What is the mechanism for the "random walk"?

   - We begin at a random page (or probabilistically, we begin at all pages with equal probability).
   - We choose one of the following:
     - We follow one of the outgoing links from this page, with probability $(1 - \alpha)$ — evenly distributed amongst all outgoing links on this page
     - We "teleport" to a page entirely at random, with probability α — evenly distributed amongst all pages (Note: if there are no out-going links, we do this with probability 1, evenly distributed.)

   (b) In terms of user behaviour, what is the significance of "teleporting"?

   - Following a link from a page is kind of obvious, based on user behaviour. "Teleporting", on the other hand, corresponds to navigating via entering a URL into the address bar (which doesn't appear to be related the content of this page).

t	$\pi(d_{(1,t)})$	$\pi(d_{(2,t)})$
0	0.5	0.5
1	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.5 = 0.3$	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.8 + 0.5 \times 0.5 = 0.7$
2	$0.3 \times 0.2 \times 0.5 + 0.7 \times 0.5 = 0.38$	$0.3 \times 0.2 \times 0.5 + 0.3 \times 0.8 + 0.7 \times 0.5 = 0.62$
3	$0.38 \times 0.2 \times 0.5 + 0.62 \times 0.5 = 0.348$	$0.38 \times 0.2 \times 0.5 + 0.38 \times 0.8 + 0.62 \times 0.5 = 0.652$

(c) The lecture example of the PageRank algorithm was given as follows:

i. Draw the implied graph corresponding to this network (which is a little smaller than the World Wide Web!).

- There are two nodes in the graph: one for each document; there is a directed edge from document 1 to document 2 (indicating a hypertext link).

ii. Which terms represent "following a link" and which represent "teleporting"?

- d_1 can only be reached via "teleporting" (there are no in-links), both terms in the sum correspond to that (the first from d_1 itself, the second from d_2)
- d_2 can be reached by following the only link from d_1, which is given by the second of the three terms in the d_2 calculations; the other two terms are "teleporting" (note that they are the same as the d_1 terms)

iii. What is the value of α in the above example? Re-do the above with $\alpha = 0.5$

- We know that following a link from document m to document n gives a term like the following:

$$d_{(n,t+1)} \quad = \quad d_{(m,t)} \times (1 - \alpha) \times \frac{1}{|m|}$$

  where $|m|$ represents the number of links in document m.
- In this case, d_1 only has one outgoing link, so we can see that d_2's score is being updated by the weight of d_1 in the previous iteration, times $(1 - \alpha)$, which means that, in this case $\alpha = 0.2$
- Changing to $\alpha = 0.5$ gives us:

t	$\pi(d_{(1,t)})$	$\pi(d_{(2,t)})$
0	0.5	0.5
1	$0.5 \times 0.5 \times 0.5 + 0.5 \times 0.5 = 0.375$	$0.5 \times 0.5 \times 0.5 + 0.5 \times 0.5 + 0.5 \times 0.5 = 0.625$
2	$0.375 \times 0.5 \times 0.5 + 0.625 \times 0.5 \approx 0.406$	$0.375 \times 0.5 \times 0.5 + 0.375 \times 0.5 + 0.625 \times 0.5 \approx 0.594$
3	$0.406 \times 0.5 \times 0.5 + 0.594 \times 0.5 \approx 0.398$	$0.406 \times 0.5 \times 0.5 + 0.406 \times 0.5 + 0.594 \times 0.5 \approx 0.602$